



# Genome assembly of a Mesoamerican derived variety of lima bean: a foundational cultivar in the Mid-Atlantic USA

Randall J. Wisser <sup>1,2,\*†</sup>, Sara J. Oppenheim,<sup>3,†</sup> Emmalea G. Ernest,<sup>4</sup> Terence T. Mhora,<sup>1,‡</sup> Michael D. Dumas,<sup>1</sup> Nancy F. Gregory,<sup>1</sup> Thomas A. Evans,<sup>1</sup> and Nicole M. Donofrio <sup>1,\*</sup>

<sup>1</sup>Department of Plant and Soil Sciences, University of Delaware, Newark, DE 19716, USA

<sup>2</sup>Laboratoire d'Ecophysiologie des Plantes sous Stress Environnementaux, INRAE, Univ. Montpellier, SupAgro, 34060 Montpellier, France

<sup>3</sup>Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA

<sup>4</sup>Cooperative Extension, University of Delaware, Georgetown, DE 19947, USA

\*Corresponding authors: Department of Plant and Soil Sciences, University of Delaware, 152 Townsend Hall, 531 S. College Avenue, Newark, DE 19716, USA.

Email: ndonof@udel.edu (N.M.D.); INRAE LEPSE, Bâtiment 7, 2 Place Pierre Viala, 34060 Montpellier, France. Email: randall.wisser@inrae.fr (R.J.W.)

<sup>†</sup>These authors contributed equally to this study.

<sup>‡</sup>Present address: FMC Stine Research Center, 1090 Elkton Road, Newark, DE 19711, USA.

## Abstract

Lima bean, *Phaseolus lunatus*, is closely related to common bean and is high in fiber and protein, with a low glycemic index. Lima bean is widely grown in the state of Delaware, where late summer and early fall weather are conducive to pod production. The same weather conditions also promote diseases such as pod rot and downy mildew, the latter of which has caused previous epidemics. A better understanding of the genes underlying resistance to this and other pathogens is needed to keep this industry thriving in the region. Our current study sought to sequence, assemble, and annotate a commercially available cultivar called Bridgeton, which could then serve as a reference genome, a basis of comparison to other *Phaseolus* taxa, and a resource for the identification of potential resistance genes. Combined efforts of sequencing, linkage, and comparative analysis resulted in a 623 Mb annotated assembly for lima bean, as well as a better understanding of an evolutionarily dynamic resistance locus in legumes.

**Keywords:** Lima bean; common bean; resistance genes; partial resistance

## Introduction

Lima bean (*Phaseolus lunatus* L.) was independently domesticated in Mesoamerica and the Andes (Motta-Aldana *et al.* 2010; Serrano-Serrano *et al.* 2012) and is now cultivated throughout the world. Lima bean is high in fiber, protein, and slow-release carbohydrates, making it a healthy, low glycemic index food (Bello-Pérez *et al.* 2007).

In the United States, lima bean is grown predominantly in the Mid-Atlantic, specifically New Jersey, Delaware, and Maryland. In Delaware alone, lima bean production is approximately a \$9.8 million industry (USDA-NASS, 2017). The same conditions that are conducive to robust pod production, however, are also conducive to the development and proliferation of several oomycete pathogens, including pod rot caused by *Phytophthora capsici* (Davidson *et al.* 2002, 2008; Evans *et al.* 2007) and downy mildew caused by *Phytophthora phaseoli*. In 1998, a new race of *P. phaseoli* emerged that decimated lima bean production in Delaware, prompting breeding for resistance genes and increased studies on this important plant, and its pathogens (Evans *et al.* 2002, 2007).

In 2016, Mhora *et al.* used bulked segregant analysis (BSA) to map a resistance locus effective against the predominant field race F of *P. phaseoli*. To our knowledge, no other resistance loci have been mapped in lima bean. Collinearity analysis with the *P. vulgaris* (common bean) genome revealed homology with a resistance gene (R-gene) dense

region containing different subtypes of nucleotide-binding site leucine-rich repeat (NLR) genes (Schmutz *et al.* 2014). Without a reference genome for *P. lunatus*, Mhora *et al.* were unable to compare the gene content in lima bean. Thus, the limited genomic resources for lima bean have hindered further dissection of this region of the genome and genome-wide analysis of disease resistance.

In addition to R-gene mediated defense, research into lima bean diversity and the genetic basis of more complex traits will require comprehensive genomic data resources. Therefore, we sequenced the *P. lunatus* cultivar Bridgeton, a variety of Mesoamerican origin that was favored by farmers in the Mid-Atlantic USA until the emergence of *P. phaseoli* race F. Furthermore, we generated linkage and syntenic maps to form larger scaffolds, and we performed QTL analysis of a slow mildewing phenotype that provides partial resistance against *P. phaseoli* (Santamaria *et al.* 2018). As a new reference genome of lima bean, this report provides a foundation for future work.

## Materials and methods

### Biological materials

A single seed of Bridgeton, namely "Bridgeton-DES4" (reproduced from NPGS PI 549508), was chosen to self-propagate seed for the

Received: April 15, 2021. Accepted: May 25, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

reference genome. Progeny of Bridgeton-DES4 was grown in continuous darkness to generate etiolated tissue for DNA extraction. Separate extractions were performed on tissue from individual plants using the Qiagen Maxi DNA extraction kit, followed by a DNA purification (Greco et al. 2014). The DNA was checked for purity using the NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, MA, USA) and quantified using both Picogreen (Thermo Fisher Scientific, MA, USA) and QUBIT (Thermo Fisher Scientific, MA, USA). Fragment analysis and agarose gel electrophoresis were used to confirm the presence of nonfragmented, high molecular weight DNA.

## Sequencing

The DNA from a single, high-quality Bridgeton-DES4 plant extract was shipped to NRGene (San Diego, CA, USA) and subjected to library construction according to their protocols. Five size fractions were selected ranging from 470 bp to 10 kb to construct sequencing libraries following the manufacturer's protocols (Illumina, San Diego, CA, USA). The TruSeq DNA Sample Preparation Kit version 2 with no PCR amplification (PCR-free) was used to make replicate paired-end libraries for the 470 and 800 bp size fractions. The Nextera MP Sample Preparation Kit was used to make mate-pair (MP) libraries with 2–5, 5–7, and 7–10 kb jumps. The 470-bp libraries were sequenced as  $2 \times 265$  nucleotides on the HiSeq2500 v2 in rapid mode. The 800-bp libraries and part of the three MP libraries were sequenced as  $2 \times 160$  bp nucleotides on the HiSeq2500 (v4 Illumina chemistry) while the remainder of the MP libraries were also sequenced as  $2 \times 150$  bp on the HiSeq4000. A total of 246 Gb sequencing data (equivalent to  $\sim 360X$  genomic coverage, based on an estimated genome size of 686 Mb). All library construction and sequencing were performed at Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign.

## Assembly

Genome assembly was conducted using the DeNovoMAGIC™ software platform (NRGene, Ness Ziona, Israel). This is a De Bruijn graph-based assembler, designed to efficiently extract the underlying information in the raw reads to solve the complexity of the De Bruijn graph due to genome polyploidy, heterozygosity, and repetitiveness. This task is accomplished using accurate-reads-based traveling in the graph that iteratively connects consecutive phased contigs over local repeats to generate long phased scaffolds (Lu et al. 2015; Hirsch et al. 2016; Avni et al. 2017; Zimin et al. 2017; Zhao et al. 2017).

In brief, the algorithm included the following steps:

- 1) Preprocessing: PCR duplicates, Illumina adaptor AGATCGGAAGAGC, and Nextera linkers (for MP libraries) were removed. The PE 450 bp  $2 \times 265$  bp libraries overlapping reads were merged with minimal required overlap of 10 bp to create the stitched reads.
- 2) Error correction: Following preprocessing, merged PE reads were scanned to detect and filter reads with putative sequencing error (contain a subsequence that does not reappear several times in other reads).
- 3) Contigs assembly: The first step of the assembly consists of building a De Bruijn graph (kmer = 127 bp) of contigs from all of the PE and MP reads. Next, PE reads were used to find reliable paths in the graph between contigs to resolve repeats and extend the contigs.
- 4) Scaffolds assembly: Contigs were linked into scaffolds with PE and MP information, estimating gaps between the contigs according to the expected distance of PE and MP links.
- 5) Fill Gaps: A final gap-filling step used PE and MP links and De Bruijn graph information to detect a unique path connecting the gap edges.

## Linkage and syntenic maps

Linkage and synteny mapping approaches were used to cluster and order the assembled scaffolds into draft pseudomolecules. Linkage mapping was performed using genotyping-by-sequencing (GBS) data on 163 F<sub>2</sub> progeny from a cross between cultivars Cypress and Jackson Wonder. 192-plex sequencing libraries were constructed following the protocol by Manching et al. (2017) using RASP-2.0 adapters redesigned for Csp6I/MspI and Csp6I/Taq<sup>I</sup> pairs of restriction enzymes. Samples included the F<sub>2</sub> progeny along with replicate samples of the parents and F<sub>1</sub> plants as well as a negative control with no DNA. Separate libraries were constructed for Csp6I/MspI and Csp6I/Taq<sup>I</sup>. Sequencing was performed at the University of Delaware's Sequencing and Genotyping Center on a HiSeq2500 run in rapid mode at  $1 \times 151$  bp.

Processing of GBS data was performed using RedRep (<https://github.com/UD-CBCB/RedRep>). First, FASTQ files from two lanes of sequencing were merged for the corresponding libraries. Following quality control and barcode deconvolution, FASTQ files from the same barcode were merged and then mapped to the NRGene sequence assembly. Variants were typed using HaplotypeCaller, and the genotype matrix was filtered as follows. VCFtools (Danecek et al. 2011) was used to set genotype calls to missing if fewer than three reads supported the call. The resulting genotype matrix was processed in R version 3.4.1 (R Core Team 2018) with custom scripts to filter markers that (1) had greater than 75% missing data; (2) had inconsistent genotype calls between parental replicates; (3) were heterozygous in either parent; (4) did not have the expected parent-hybrid trio genotypes; and (5) had F<sub>2</sub> allele frequencies less than 15% or greater than 85%.

Using sequencing scaffolds containing at least five markers, missing data were imputed using LB-Impute (Fragoso et al. 2016). A linkage map was constructed with QTL IciMapping software v 4.1.0.0 (Meng et al. 2015) [run settings: DIS (20 cM) grouping function; RECORD ordering algorithm; SARF (window size = 5) rippling criterion].

Chromosomer (Tamazian et al. 2016) was used to align the NRGene scaffolds against the *P. vulgaris* reference genome (Schmutz et al. 2014) downloaded from JGI ("*Pvulgaris\_442\_v2.0.softmasked.fa.gz*"). First, while retaining softmasked sequences in the reference genome identified by analysis with RepeatMasker, LAST (Kielbasa et al. 2011) was used to identify and softmask additional repeat sequences using the NEAR seeding scheme (run settings: -uNEAR and -R11). Following guidelines for human-ape alignments (<https://github.com/mcfrith/last-genome-alignments>, last accessed December 2019), substitution and gap frequencies were determined with last-train (run settings: -revsym -matsym -gapsym -E0.05 -C2) and lastal alignment was performed (run settings: -fMAF -K2 -m50 -E0.05 -C2, and -p corresponded to the output from last-train). Alignments between repeat sequences were discarded with last-postmask and a python script was used to retain only the top two LAST matches per query sequence. The resulting output was used to run fragmentmap and assembler routines of Chromosomer. Finally, LAST was used to align the Chromosomer assembled *P. lunatus*

genome to the *P. vulgaris* reference genome (run settings: -m50 -E0.05 -C2).

## QTL analysis

Disease reaction to race F of downy mildew (caused by *P. phaseoli*) was determined for 384 F<sub>2:3</sub> families of the Cypress (resistant; slow mildewing phenotype) X Jackson Wonder (susceptible) cross (163 of the F<sub>2</sub> parents of these families were used to construct the linkage map; see above). Plants were grown in a greenhouse humidity chamber and inoculated at emergence as described by Santamaria et al. (2018). All families were replicated across sequential plantings (four batches across time) with five plants per family in a single pot in each round. Pots were arranged in a randomized incomplete block design with six subblocks augmented with repeated checks of resistant, tolerant, and susceptible varieties. Individual plants were measured for lesion length on the stem and rated for the quantity of sporulation on a 1–5 scale. To account for differences in stem length, lesion length was standardized by plant height. Adjusted means for sporulation rating and height-standardized lesion length on the F<sub>2:3</sub> progeny was used as an estimate of the corresponding F<sub>2</sub> parent phenotype for QTL analysis. QTL IciMapping software v 4.1.0.0 (Meng et al. 2015) was used to perform inclusive composite interval mapping with an LOD threshold of 2.5 and step size of 1 cM.

## Gene annotation

Prior to structural annotation, repetitive elements were identified and masked (Supplementary File S1) using RepeatMasker (Smit et al. 2013) with a custom library that included all Viridiplantae entries from Repbase (Bao et al. 2015) along with repetitive elements from *P. vulgaris* (Gao et al. 2014). Using the masked assembly, gene predictions were generated by AUGUSTUS with the *Arabidopsis thaliana* training set (Keller et al. 2011; Scalzitti et al. 2020). The completeness of the genome assembly and AUGUSTUS gene models was analyzed with BUSCO, which measures completeness in terms of evolutionarily informed expectations of gene content (Simão et al. 2015). The BUSCO Embryophyta dataset (1440 single-copy conserved genes) was used as a reference.

Annotation of the predicted genes was undertaken with multiple tools. First, BLASTp searches (-evalue 1e-10) were performed against a custom BLAST database that included all Viridiplantae sequences from the NCBI nr database (Wheeler et al. 2008). For sequences that had no hit, a second BLASTp search was performed against all sequences in the nr database. There were 14 RNA-Seq libraries in PRJNA596114 (*P. lunatus* genome), generated from pods (12 libraries), leaves (1 library), and flowers (1 library). We used MagicBlast at NCBI to map these reads to the predicted genes. The resulting SAM file was converted to BAM format and indexed, followed by samtools idxstats to identify predicted genes with either no transcript coverage or those that were covered over their full length. For the remaining genes with partial coverage, bedtools genomecov was used to calculate the fraction of transcript coverage across each gene. In addition, using BLASTp (-evalue 1e-10), the predicted proteins for *P. lunatus* were compared to the *P. vulgaris* reference proteome (<https://www.uniprot.org/proteomes/UP000000226>, last accessed July 2020).

We used protein domain analysis to annotate NLR genes. To support comparison with common bean, we followed the same procedure for annotating NLRs described by Schmutz et al. (2014).

## Results and discussion

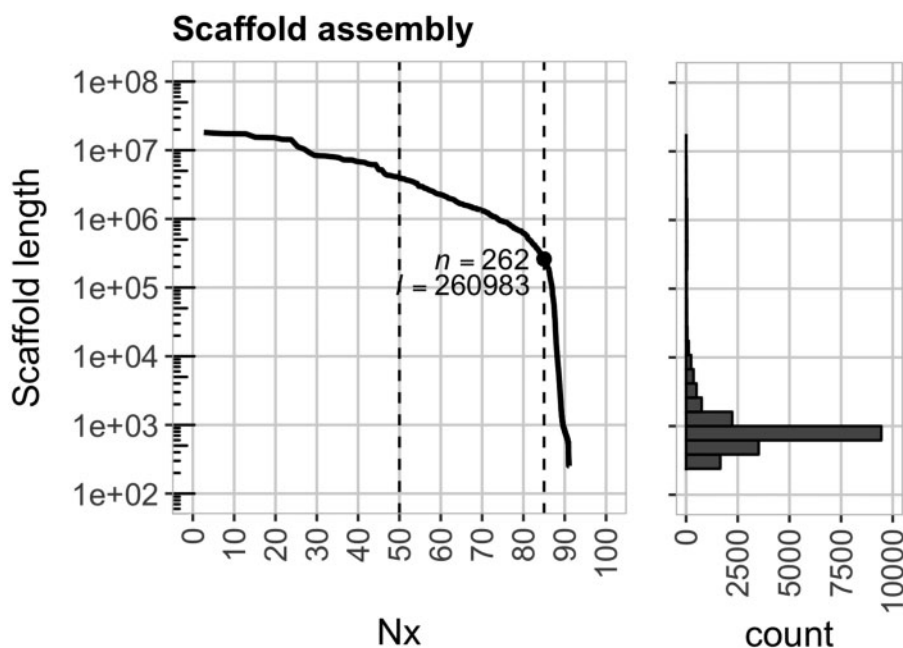
We report a 623 Mb genome assembly of the Mesoamerican lima bean cultivar Bridgeton. This constitutes ~91% of the 686 Mb genome estimated by flow cytometry for *P. lunatus* (Pellicer and Leitch 2020). The final assembly has an average depth of ~70X (Supplementary Table S1) and is comprised of 19,316 scaffolds, 36 and 262 of which captured 50% and 85% of the expected genome size, respectively (Figure 1), with a corresponding N50 of 3.99 Mb and N85 of 0.26 M. The N50 was 5.17 Mb with respect to the sequence space (as opposed to the expected genome size). Consistent with the nucleotide composition of higher plant genomes, the GC content for lima bean was 38%, the same as its closest relative with a reference genome, common bean (Schmutz et al. 2014). The genome was analyzed for completeness, returning a BUSCO score of 93% (Table 1). Although this represents a nearly complete genome assembly for lima bean, the assembly is comprised of thousands of scaffolds that vastly exceed the 11 (2n = 22) chromosomes of *P. lunatus* (Bonifácio et al. 2012).

Linkage and synteny mapping methods were used to further tether the assembly scaffolds. Using GBS data, a linkage map was constructed from 942 markers present in 46 scaffolds that captured ~50% of the sequence space. The largest 11 linkage groups constituted a 670 cM map (see MAP directory in Supplementary File S2) comprised of 898 markers among 41 scaffolds (25 of which were in the 36 N50 set and all of which were within the N85 set) that together captured 49% of the sequence space. In every case, markers within a given scaffold mapped to a single linkage group. However, within LGs 1, 3, and 5, markers were ordered on the genetic map such that sections of different scaffolds interleaved with other scaffolds (Supplementary Table S2).

Syntenic analysis with the common bean genome (Schmutz et al. 2014) ordered 3839 of the *P. lunatus* scaffolds, which were distributed among all 11 *P. vulgaris* chromosomes (Supplementary File S3). These corresponded to many different scaffolds in the *P. lunatus* assembly (~230 Mb or 37% of the assembled genome) than those anchored to the genetic map. There were 16 scaffolds anchored by both maps which comprised ~105 Mb. Together, nearly 70% of the sequence space was ordered by genetic mapping or synteny analysis, but with only 25% intersection, these separate maps could not be well integrated. Additional work will be required to construct chromosome-scale pseudomolecules. Nevertheless, as described below, combining all of the map data helped to identify a homologous section of the *P. lunatus* and *P. vulgaris* genomes associated with variation in disease resistance.

We mapped a QTL on LG 9 (*PlPp\_LG9.1*) that explained ~50% of the phenotypic variation in partial resistance to *P. phaseoli* (see BIP directory in Supplementary File S2), a slow downy-mildewing phenotype. Previously, with a different population, Mhora et al. (2016) used BSA to map a race-specific major effect, race-specific gene also associated with resistance to *P. phaseoli*. Anchoring the marker sequences from QTL and BSA mapping onto our genome assembly showed that both of these resistance loci reside at the same region of the genome, indicating they are linked or that *PlPp\_LG9.1* is a weak allele of the race-specific resistance gene.

Markers present in scaffold 25,456 were associated by both QTL and BSA mapping. Two additional scaffolds contained BSA-associated markers, but these scaffolds were absent from the linkage map used for QTL analysis [due to marker QC filtering (scaffold 3610) and failure to join a linkage group (scaffold 974); scaffold 974 contained markers with the strongest BSA



**Figure 1** Summary statistics for the assembly of lima bean cultivar Bridgeton. The plot on the left shows the scaffold length as a function of the contiguity value (Nx). The N50 and N85 are marked by the dashed vertical lines where the corresponding number (n) and cumulative length (l) of scaffolds are noted. The marginal histogram plot on the right shows the number of scaffolds at different lengths.

**Table 1** Genome statistics and annotated gene data

Total scaffolds	19,316
Total contigs	58,167
Complete BUSCOs	1,341 (93%)
Scaffolds with predicted proteins	3,417
Protein-coding genes	64,541
Genes with transcript evidence	40,308
Proteins with <i>P. vulgaris</i> hit	32,505
Proteins with other Viridiplantae hit	1,356
Proteins with non-Viridiplantae hit	2,363
Proteins with no hit in nr database	28,317

association]. Synteny mapping helped to delimit the physical section of the genome containing both resistance loci, but this required a lowering the standard sensitivity threshold for chromosomeric ( $-r$  1.01; default is 1.2). Based on these results, 15 scaffolds spanned the QTL and BSA-associated loci with flanking markers in scaffolds 3610 and 25,456 (see Supplementary File S4 fragment map). Consistent with previous findings (Mhoro et al. 2016), this corresponded to a homologous section on chromosome 4 in *P. vulgaris* (coordinates: 75,772–1,552,787). The approximate length of *P. lunatus* scaffold sequences between the flanking markers was 1.3 Mb, smaller than *P. vulgaris* by ~175 kb.

### Genome-wide annotation

Augustus predicted 64,541 genes from the masked assembly (Supplementary File S5). As is typical of draft genome assemblies, this is likely an overestimation of the true gene number in *P. lunatus*. Inflated gene counts can result from assembly fragmentation (a single gene sequence spread across multiple contigs) and failure to join distant exons together in a single transcript (Denton et al. 2014). The BUSCO analysis also suggests overestimation of the true gene number: ~13% of the core genes were either duplicated or fragmented.

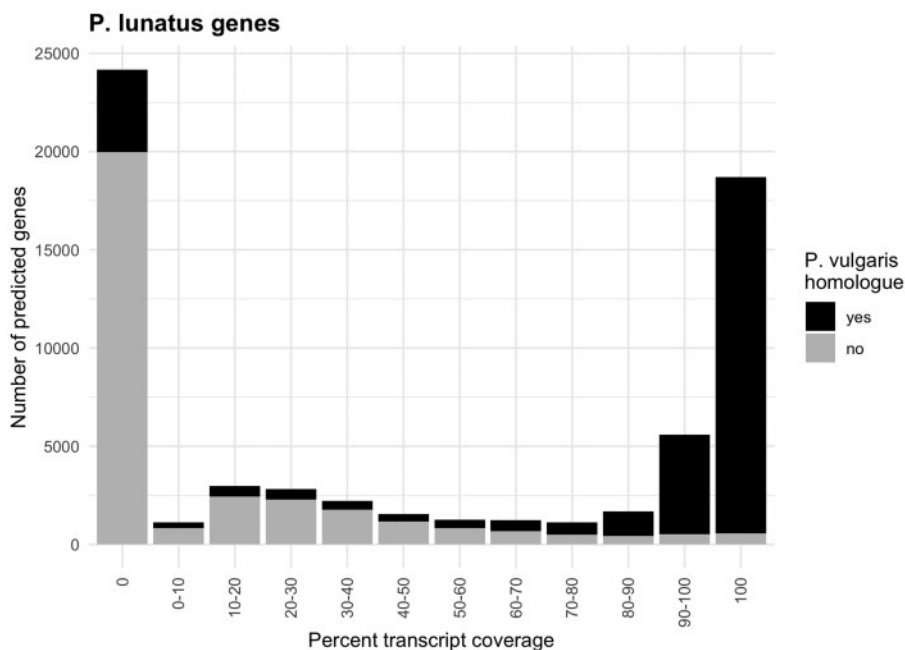
We compared the predicted proteins for *P. lunatus* to protein annotations for *P. vulgaris*, which diverged from *P. lunatus* ~4 mya

(Delgado-Salinas et al. 2006; Bitocchi et al. 2017). The *P. vulgaris* genome has an estimated size of 587 Mb, with 98% of the sequence anchored to 11 pseudomolecules that contain 27,197 protein-coding genes supported by transcripts (Schmutz et al. 2014). We found that ~50% (32,505) of the *P. lunatus* proteins had hits to *P. vulgaris*, which encompassed 96% of the *P. vulgaris* proteins (Supplementary File S6). Using public transcriptome data for lima bean, we found that the vast majority of *P. vulgaris* homologs had transcript evidence in *P. lunatus* (Figure 2; Supplementary File S7), while most, but not all, of the nonhomologous genes were likely to be errors in *de novo* gene prediction. Of the 32,036 predicted *P. lunatus* proteins without a match in *P. vulgaris*, 28,317 proteins had no hits to any sequence in the nr database and tended to be shorter (mean of 237 aa) than those with hits (mean of 514 aa). However, 1356 had hits to other Viridiplantae species, and another subset of 2363 had hits to non-Viridiplantae entries in the nr database that were predominantly bacterial sequences. There were 903 hits to Enterobacteriaceae, which are the dominant members of many phyllosphere bacterial communities (Cernava et al. 2019), 359 hits to Rhizobiaceae, a family of plant-associated bacteria (Spaink et al. 2012), and 254 hits to Flavobacteria, which are thought to contribute to plant growth and protection (Kolton et al. 2016). Only 132 of the non-Viridiplantae nr hits were to eukaryotes (90 metazoa, 31 fungi, 6 Alveolata, 2 Stramenopiles, 2 Euglenozoa, and 1 Pyrenomonadales).

### Resistance genes

One of our objectives in sequencing the lima bean genome was to catalog R-genes commonly associated with disease resistance in plants. The majority of these are NLR genes defined by two major domains: a centrally located nucleotide-binding site domain, which has ATPase activity (NB-ARC) and a C-terminus leucine-rich repeat domain (LRR) (van Ooijen et al. 2008; Takken and Govere 2012), but additional non-NLR types with coiled-coil domains (CNS and CNLs) are also important. The *P. vulgaris* genome contained 376 such R-genes (Schmutz et al. 2014), while the





**Figure 2** Transcript and comparative assessment of predicted genes in the Bridgeton genome. A histogram of the percent transcript coverage for predicted genes in the Bridgeton assembly. The fraction of encoded protein homologs in the common bean is indicated per class.

*P. lunatus* Bridgeton assembly contained 190, with the major difference due to a paucity of NLRs with a Toll/Interleukin Receptor-1 domain in lima bean (Supplementary File S8). The *P. vulgaris* genome has three particularly large clusters containing some 40R-genes at the ends of chromosomes 4, 10, and 11 (Richard et al. 2018). The chromosome 4 cluster, referred to as the B4 locus, is homologous with the locus described in this study that was associated with partial and complete resistance to downy mildew. The B4 locus in common bean has been noted for ectopic recombination resulting in the accumulation of CNLs on chromosome 4. The homologous section in lima bean contained many fewer R-genes (9 compared to 39 in the common bean), but six of these, all of which were located on scaffold 974, contained a coiled-coil domain (Supplementary File S8).

## Conclusions

Closing a major gap in resources for lima bean, this study reports the reference genome for a *P. lunatus* cultivar that was foundational to lima bean production in the east coast. *De novo* protein predictions showing high similarity to 96% of the encoded genes for *P. vulgaris* (common bean) corresponded to 32,505 genes in lima bean, most of which were supported by transcriptome data (Figure 2). The sequenced variety, Bridgeton, was the primary founder of modern cultivars for the Mid-Atlantic. In this region of the USA, diseases limit the production of lima bean. Using the genome assembly, we consolidated genetic map data for loci associated with partial, race nonspecific resistance and complete, race-F-specific resistance to *P. phaseoli*, the causal agent of downy mildew. These loci colocalized in a segment that aligns to a section of chromosome 4 of common bean which is enriched with canonical R-genes and genetic associations with resistance to different diseases—the B4 R-gene cluster (David et al. 2008, 2009). Despite many fewer R-genes at the homologous locus in the Bridgeton genome, the presence of coiled-coil type R-genes is a shared feature. This is consistent with prior findings of ectopic recombination events at the B4 locus that predates the divergence

of lima and common bean (David et al. 2009). Thus, the reference genome of lima bean enabled the identification of a putative hotspot for the evolution of resistance alleles, which merits further research. This report provides a new genomic resource for investigations into the diversity and evolution of legumes.

## Data Availability

The genome sequence data for this study is available under the NCBI BioProject PRJNA647124, for BioSample SAMN15394833 (*P. lunatus* cv. Bridgeton). Raw read data from the individual libraries at the Sequence Read Archive include: SRX9040258, SRX9040257, SRX9040256, SRX9040255, SRX9040254, SRX9040253, SRX9040252, SRX9040251. Supplementary material is available at figshare: <https://doi.org/10.25387/g3.14398910>. This Supplementary material includes eight files: a repeat masked version of the genome assembly (Supplementary File S1); the linkage map and QTL mapping results from IciMapping (Supplementary File S2); synteny maps from Chromosomer (Supplementary Files S3 and S4); all genes predicted by Augustus (Supplementary File S5); proteins of *P. lunatus* with a match to a *P. vulgaris* protein (Supplementary File S6); transcript coverage for *P. lunatus* predicted genes (Supplementary File S7); and *P. lunatus* R-gene annotations and comparison with *P. vulgaris* (Supplementary File S8).

## Acknowledgments

The authors thank Colin Scanlon for his assistance with plant growth and phenotyping, and Dr Karol Miaskiewicz for facilitating computational activities on BIOMIX.

## Funding

The authors gratefully acknowledge funds from the Delaware Department of Agriculture, grant award number 58-8042-7-079 to N.M.D., T.T.M., and T.A.E., and 15-SCBGP-DE-0028 to E.G.E. and

R.J.W. Some of this work was performed on the BIOMIX compute cluster, made possible through funding from Delaware INBRE (National Institutes of Health, National Institute of General Medical Sciences P20GM103446), the State of Delaware, and the Delaware Biotechnology Institute.

## Conflicts of interest

The authors have no conflicts of interest to declare.

## Literature cited

- Avni R, Nave M, Barad O, Baruch K, Twardziok SO, et al. 2017. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*. 357:93–97.
- Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 6:11.
- Bello-Pérez LA, Sáyago-Ayerdi SG, Chávez-Murillo CE, Agama-Acevedo E, Tovar J. 2007. Proximal composition and in vitro digestibility of starch in lima bean (*Phaseolus lunatus*) varieties. *J Sci Food Agric*. 87:2570–2575.
- Bitocchi E, Domenico R, Bellucci E, Rodriguez M, Murgia M, et al. 2017. Beans (*Phaseolus* spp.) as a model for understanding crop evolution. *Front Plant Sci*. 8:722.
- Bonifácio EM, Fonsêca A, Almeida C, Dos Santos KG, Pedrosa-Harand A. 2012. Comparative cytogenetic mapping between the lima bean (*Phaseolus lunatus* L.) and the common bean (*P. vulgaris* L.). *Theor Appl Genet*. 124:1513–1520.
- Cernava T, Erlacher A, Soh J, Sensen CW, Grube M, et al. 2019. Enterobacteriaceae dominate the core microbiome and contribute to the resistome of arugula (*Eruca sativa* Mill.). *Microbiome*. 7:13.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al.; 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics*. 27:2156–2158.
- David P, Sévignac M, Thareau V, Catillon Y, Kami J, et al. 2008. BAC end sequences corresponding to the B4 resistance gene cluster in common bean: a resource for markers and synteny analyses. *Mol Genet Genomics*. 280:521–533.
- David P, Chen NW, Pedrosa-Harand A, Thareau V, Sévignac M, et al. 2009. A nomadic subtelomeric disease resistance gene cluster in common bean. *Plant Physiol*. 151:1048–1065.
- Davidson CR, Mulrooney RP, Carroll RB, Evans TA. 2002. First report of *Phytophthora capsici* on lima bean in Delaware. *Plant Disease* 85:886.
- Davidson CR, Evans TA, Mulrooney RP, Gregory NF, Carroll RB, et al. 2008. Lima Bean Downy Mildew Epiphytotic Caused by New Physiological Races of *Phytophthora phaseoli*. *Plant Dis*. 92:670–674.
- Delgado-Salinas A, Bibler R, Lavin M. 2006. Phylogeny of the genus *Phaseolus* (Leguminosae): a recent diversification in an ancient landscape. *Systematic Botany*. 31:779–791.
- Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, et al. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol*. 10:e1003998.
- Evans TA, Davidson CR, Dominiak JD, Mulrooney RP, Carroll RB, Antonius SH. 2002. Two new races of *Phytophthora phaseoli* from lima bean in Delaware. *Plant Disease*. 86:813.
- Evans TA, Mulrooney RP, Gregory NF, Kee E. 2007. Lima bean downy mildew: impact etiology and management strategies for Delaware and the mid-Atlantic U.S. *Plant Dis*. 91:128–135.
- Fragoso CA, Heffelfinger C, Zhao H, Dellaporta SL. 2016. Imputing genotypes in biallelic populations from low-coverage sequence data. *Genetics*. 202:487–495.
- Gao D, Abernathy B, Rohksar D, Schmutz J, Jackson SA. 2014. Annotation and sequence diversity of transposable elements in common bean (*Phaseolus vulgaris*). *Front Plant Sci*. 5:339.
- Greco M, Sáez CA, Brown MT, Bitonti MB. 2014. A simple and effective method for high quality co-extraction of genomic DNA and total RNA from low biomass *Ectocarpus siliculosus*, the model brown alga. *PLoS One*. 9:e96470.
- Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, et al. 2016. Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *Plant Cell*. 28:2700–2714.
- Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*. 27:757–763.
- Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res*. 21:487–493.
- Kolton M, Erlacher A, Berg G, Cytryn E. 2016. The flavobacterium genus in the plant holobiont: ecological, physiological, and applicative insights. In: S Castro-Sowinski, editor. *Microbial Models: From Environmental to Industrial Sustainability*. Microorganisms for Sustainability. Singapore: Springer. pp 189–207.
- Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, et al. 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat Commun*. 6:6914.
- Manching H, Sengupta S, Hopper KR, Polson SW, Ji Y, et al. 2017. Phased genotyping-by-sequencing enhances analysis of genetic diversity and reveals divergent copy number variants in maize. *G3 (Bethesda)*. 7:2161–2170.
- Meng L, Li H, Zhang L, Wang J. 2015. QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J*. 3:269–283.
- Mhora TT, Ernest EG, Wissner RJ, Evans TA, Patzoldt ME, et al. 2016. Genotyping-by-sequencing to predict resistance to lima bean downy mildew in a diversity panel. *Phytopathology*. 106:1152–1158.
- Motta-Aldana JR, Serrano-Serrano ML, Hernández-Torres J, Castillo-Villamizar G, Debouck DG, et al. 2010. Multiple origins of lima bean landraces in the Americas: evidence from chloroplast and nuclear DNA polymorphisms. *Crop Sci*. 50:1773–1787.
- Pellicer J, Leitch IJ. 2020. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol*. 226:301–305.
- R Core Team. 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/> (Accessed: 2020 May).
- Richard MMS, Gratias A, Thareau V, Kim KD, Balzergue S, et al. 2018. Genomic and epigenomic immunity in common bean: the unusual features of NB-LRR gene family. *DNA Res*. 25:161–172.
- Santamaria L, Ernest EG, Gregory NF, Evans TA. 2018. Inheritance of resistance in lima bean to *Phytophthora phaseoli*, the causal agent of downy mildew of lima bean. *Am Soc Hortic Sci*. 53:777–781.
- Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. 2020. A benchmark study of *ab initio* gene prediction methods in diverse eukaryotic organisms. *BMC Genomics*. 21:293.
- Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, et al. 2014. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat Genet*. 46:707–713.
- Serrano-Serrano ML, Andueza-Noh RH, Martínez-Castillo J, Debouck DG, Chacón MI. 2012. Evolution and domestication of lima bean in Mexico: evidence from ribosomal DNA. *Crop Sci*. 52:1698–1712. doi:10.2135/cropsci2011.12.0642.

- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31:3210–3212.
- Smit AFA, Hubley R, Green P. 2013. 2015 RepeatMasker Open-4.0. <http://www.repeatmasker.org> (Accessed: 2020 April).
- Spaink HP, Kondorosi A, Hooykaas PJ. 2012. *The Rhizobiaceae: Molecular Biology of Model Plant-Associated Bacteria*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Takken FL, Goverse A. 2012. How to build a pathogen detector: structural basis of NB-LRR function. *Curr Opin Plant Biol*. 15:375–384.
- Tamazian G, Dobrynin P, Krasheninnikova K, Komissarov A, Koepfli KP, et al. 2016. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. *Gigascience*. 5:11.
- USDA-National Agricultural Statistics Service (USDA-NASS). 2017. Census of Agriculture. <https://www.nass.usda.gov/Publications/AgCensus/2017/index.php> (Accessed: 2021 February).
- van Ooijen G, Mayr G, Kasiem MM, Albrecht M, Cornelissen BJ, Takken FL. 2008. Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. *J Exp Bot*. 59:1383–1397.
- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 36:D13–D21.
- Zhao G, Zou C, Li K, Wang K, Li T, et al. 2017. The *Aegilops tauschii* genome reveals multiple impacts of transposons. *Nat Plants*. 3: 946–955. doi:10.1038/s41477-017-0067-8.
- Zimin AV, Puiu D, Lyons E, You FM, Lu FH, et al. 2017. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature*. 551: 498–502.

Communicating editor: J. Ma